Steelmanning First: Charity, Virtue, and Public Reason as a Unified Method of Philosophical Inquiry

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

October 19, 2025

This paper argues that steelmanning—presenting an interlocutor's position in its strongest, most coherent form before critique—is not a conversational nicety but a core philosophical method. It fuses three traditions: (i) the principle of charity in interpretation (from Socratic dialectic through Quine and Davidson), (ii) virtue epistemology's emphasis on intellectual character (humility, courage, fairness), and (iii) deliberative public reason, which requires that claims be stated in terms others could in principle accept or evaluate (Rawls, Habermas). Read this way, steelmanning converts dispute into inquiry: it widens common ground, surfaces *cruxes* rather than caricatures, and restores the conditions under which reasons can actually meet. We show that the practice is normatively required by truth-conduciveness, justice in discourse, and the cultivation of intellectual virtue; we also delineate real limits (e.g., harm-amplifying restatements, performative self-contradictions, and radical incommensurability). Drawing on exemplars from Plato and Aquinas to Mill and Dennett, we provide a minimal, portable procedure for "steelmanning first," then test it across hard cases (science/religion disputes, ideology, identity-implicating claims). The result is a compact, tradition-grounded discipline: charitably reconstruct, locate the crux, propose the cheapest decisive test, then answer—a method fit for classrooms, journals, and civic life.

Keywords: steelmanning, principle of charity, virtue epistemology, public reason, Socratic dialectic, Aquinas, Mill, Davidson, Rawls, Habermas, deliberative democracy, disagreement, intellectual humility, Gadamer, Rapoport's rules.

A collaboration with GPT-5. CC4.0.

Outline

1. Thesis and Positioning
   1.1 Definition: steelmanning as strongest-faithful reconstruction
   1.2 Distinctions: charity vs. concession; best-case vs. straw/hollow man
   1.3 Payoff: dispute → inquiry; adversaries → co-inquirers

2. Genealogy of the Method
   2.1 Socratic dialectic: eliciting the best account before refutation (Plato)
   2.2 Aristotelian dialectic/endoxa and the ethics of fair statement
   2.3 Scholastic disputation: Aquinas's *Objectiones* → *Respondeo*
   2.4 Hermeneutics: Gadamer's fusion of horizons and good-faith repair
   2.5 Analytic charity: Quine/Davidson on maximizing truth and coherence
   2.6 Deliberative reason: Rawls/Habermas on justificatory reciprocity
   2.7 Modern articulation: Rapoport's rules; Dennett's formulation

3. Normative Grounds
   3.1 Epistemic: truth-conduciveness, Gricean cooperation, Aumann's theorem
   3.2 Ethical-political: due process for reasons in plural societies
   3.3 Virtue-epistemic: humility, courage, generosity, and integrity

4. Method: The Philosophical Steelman
   4.1 Interpretive charity and principle of repair (Davidson)
   4.2 Best-case reconstruction: supplying admissible premises, pruning inessentials
   4.3 Telos disclosure: naming the good the view aims to secure
   4.4 Public-reason rendering: reasons legible to non-adherents
   4.5 The "satisfaction test": endorsement by the interlocutor

5. Hard Cases and Limits
   5.1 Performative self-contradictions and minimal restatement
   5.2 Harm-amplifying speech and justified refusal to steelman
   5.3 Radical incommensurability: translation before adjudication
   5.4 Motte-and-bailey dynamics and boundary-setting

6. Exemplars (Textual Miniatures)
   6.1 Aquinas: strongest objections first, then answer
   6.2 Mill: live adversaries vs. "dead dogma"
   6.3 Dennett/Rapoport: endorsement-standard restatement

7. Applications
   7.1 Classroom disputation and seminar practice
   7.2 Editorial/journal peer review norms
   7.3 Civic forums and policy deliberation (public reason)
   7.4 Interfaith, science–religion, and identity-implicating debates
8. Objections and Replies
   8.1 “Charity enables bad ideas” → filtering by harm/performative incoherence
   8.2 “It stacks the deck for opponents” → reciprocity and symmetry
   8.3 “It’s inefficient” → crux-finding and cheapest decisive tests
9. A Compact Procedure (Philosophy-First)
   9.1 Reconstruct → Confirm → Identify crux → Propose test → Answer
   9.2 Exemplar paragraph template and satisfaction criterion
10. Conclusion: From Victory to Understanding
   10.1 Steelmanning as discipline of shared reason
   10.2 Cultivating persons and publics capable of disagreement without contempt
   10.3 Future work: metrics for discourse quality; pedagogy and institutional uptake

References

1. Thesis and Positioning

1.1 Definition: steelmanning as strongest-faithful reconstruction

Steelmanning is the disciplined act of presenting an interlocutor's position in its strongest form consistent with fidelity to that interlocutor's actual commitments. "Strongest" means: (i) logically valid or cogent structure, (ii) the best available premises that the interlocutor would endorse, (iii) the most relevant and probative evidence in its favor, and (iv) articulation at its optimal scope (neither overbroad nor trivial). "Faithful" means: (a) no additions they would reject, (b) no omissions that change meaning, (c) definitions and key terms preserved or clarified without substitution, and (d) hedges, exceptions, and value-aims stated as they intend.

Minimal norm. A reconstruction counts as a steelman only if the originator, when asked, says: *"Yes, that is my view, and it is put as well as or better than I put it."* Call this the endorsement criterion.

Rationale. The practice treats argument as a cooperative search for truth rather than a contest of status. It aligns with the principle of charity (interpret for maximal coherence) while imposing an additional demand: amplify the view up to its best defensible form without crossing into invention.

1.2 Distinctions: charity vs. concession; best-case vs. straw/hollow man

Charity vs. concession.

- Charity repairs surface defects (ambiguity, missing lemmas) to reveal the strongest coherent reading.
- Concession yields substantive ground—accepting a premise or conclusion you do not, upon reflection, endorse.
  Steelmanning requires charity but not concession. One can present the best case and still reject it.

Best-case vs. straw man.

- Straw man: attack a weakened caricature (exaggerated scope, omitted qualifications).
- Best-case: supply the argument's most defensible premises, strongest warrants, and right scope as its proponent would accept.
  Diagnostic: if your summary makes the view easier to refute than the proponent's own version, you likely built straw.

Best-case vs. hollow man.

- Hollow man: refute a view no one holds (a conjured opponent).
- Best-case: is anchored in a specific agent/text with citably held commitments. Diagnostic: you can name the holder(s) and quote or cite where each key commitment is expressed or entailed.

Guardrails (quick tests).

1. Admissibility test: Would they sign off on each premise?
2. Scope test: Did you preserve their intended domain and exceptions?
3. Aim test: Did you name the good or value the view is meant to secure?
4. Cost test: Did you include the most serious objections they themselves acknowledge?

1.3 Payoff: dispute → inquiry; adversaries → co-inquirers

Epistemic payoff (truth-conduciveness).
Steelmanning moves discourse from competitive dispute to cooperative inquiry. By maximizing the opposing position's strength, you reduce false positives in refutation (defeating only the weakest version) and expose the crux—the narrow hinge on which disagreement actually turns (a contested fact, a methodological norm, or a value priority). Once the crux is visible, participants can propose the cheapest decisive test (an observation, a domain restriction, a conditional policy) rather than trading intuitions.

Ethical-political payoff (justice in discourse).
Rendering another's view at its best enacts fairness owed to persons in plural settings. It lowers defensiveness, raises perceived legitimacy, and makes reciprocity more likely: once heard charitably, interlocutors are likelier to extend the same discipline. This is not civility theater; it is procedural justice for reasons.

Character payoff (intellectual virtue).
The practice trains humility (I might be wrong), courage (I might be moved), and integrity (I do not win by misrepresentation). Over time, these virtues convert "adversaries" into co-inquirers—agents jointly responsible for locating shared standards, agreeing on tests, and living with the results.

Operational upshot (one sentence).
Steelmanning redefines success: not "I defeated you," but "We located the crux, named the tests, and clarified where and why we still part ways."

2. Genealogy of the Method

2.1 Socratic dialectic: eliciting the best account before refutation (Plato)

Socrates advances inquiry by drawing out a partner's *logos*—not to humiliate, but to clarify what the other truly means. In dialogues like *Gorgias* and *Republic*, he asks for definitions, hunts hidden premises, and strengthens the interlocutor's case by tightening terms and examples before testing them by *elenchus*. The aim is a truer articulation of the other's view; refutation, when it comes, is a by-product of shared search, not point-scoring.

2.2 Aristotelian dialectic/endoxa and the ethics of fair statement

Aristotle makes charity into method. Dialectic begins from *endoxa*—the most reputable opinions—and treats them with due care: one must state opposing positions as their champions would, then resolve tensions by distinction and demonstration (*Topics*, *Rhetoric*). The ethic is clear: misrepresenting an *endoxon* corrupts inquiry; just representation dignifies both truth and audience.

2.3 Scholastic disputation: Aquinas's *Objectiones* → *Respondeo*

Aquinas institutionalizes steelmanning: every article in the *Summa* opens with the strongest objections (*Objectiones*), sometimes better phrased than in their sources, then a brief authoritative counter-thesis (*Sed contra*), and finally the reasoned answer (*Respondeo*), which often grants much to the objections before drawing principled lines. The form teaches: state the other side at full strength, then answer with reasons that acknowledge its best insights.

2.4 Hermeneutics: Gadamer's fusion of horizons and good-faith repair

For Gadamer, understanding is not extraction but a "fusion of horizons." One must enter the other's historical-linguistic world, presuming sense and repairing surface defects to reach the question the text or speaker is *really* answering. Steelmanning here is hermeneutic charity: interpret toward intelligibility first; only then adjudicate.

2.5 Analytic charity: Quine/Davidson on maximizing truth and coherence

Quine and Davidson give the principle a rationality constraint: interpret so as to maximize truth and coherence, else communication becomes impossible. Apparent error is first apportioned to translation before to belief. Steelmanning inherits this: prefer the reading that makes an opponent minimally wrong and structurally rational, unless overwhelming evidence forces a harsher verdict.

2.6 Deliberative reason: Rawls/Habermas on justificatory reciprocity

Deliberative theories add a public norm: in a plural society, reasons must be given so others could, in principle, accept or evaluate them (Rawls's "public reason," Habermas's "communicative rationality"). Steelmanning becomes reciprocity in practice: present your interlocutor's view in the very terms that would make it most assessable to free and equal others—then ask the same in return.

2.7 Modern articulation: Rapoport's rules; Dennett's formulation

Rapoport (popularized by Dennett) turns the virtue into rules: (1) state the opponent's position so clearly and fairly that they say "Thanks, I wish I'd put it that way"; (2) list agreements; (3) mention what you learned; only then (4) rebut. This operationalizes the whole genealogy: Socratic elicitation, Aristotelian fairness, Scholastic ordering, hermeneutic repair, analytic charity, and deliberative reciprocity—condensed into a practice any disputant can adopt.

3. Normative Grounds

3.1 Epistemic: truth-conduciveness, Gricean cooperation, Aumann's theorem

Truth-conduciveness. Steelmanning reduces Type-I error in critique (defeating only a weak caricature) and increases power to detect real disagreement. By reconstructing the *best* admissible version of a view, we aim at veridical adjudication—if that version fails, the target likely fails; if it survives, we have learned where our own model is thin.

Gricean cooperation. Grice's conversational maxims (quality, quantity, relation, manner) presuppose a shared aim of informative truth-telling. Steelmanning operationalizes these maxims:

- *Quality*: do not attribute what the speaker would disown; maximize warranted content in their favor.
- *Quantity*: neither omit decisive premises nor pad with weak points.
- *Relation*: highlight the claims most relevant to the dispute's crux.
- *Manner*: clarify definitions and scope so the position is legible.
  This restores the cooperative baseline required for rational uptake.

Aumann's agreement theorem. If two Bayesian agents share (i) common priors, (ii) common knowledge of each other's posteriors, they cannot "agree to disagree." Steelmanning approximates the theorem's preconditions: by making each side's reasons and evidential weightings transparent, it narrows the gap between *apparent* disagreement (talking past one

another) and residual crux (genuine conflict in priors, values, or data). Even when common priors are unrealistic, steelmanning locates the minimal set of divergences that must be addressed for convergence, turning debate into diagnosis.

3.2 Ethical-political: due process for reasons in plural societies

Reciprocity and justification. In a polity of free and equal persons, legitimacy requires that coercive rules be justifiable to those bound by them. Steelmanning enacts this reciprocity at the micro-level: it treats interlocutors as co-authors of the justificatory space by representing their reasons as they would.

Due process for reasons. As legal due process forbids condemning a person unheard, discursive due process forbids dismissing a view misdescribed. Steelmanning is the procedural justice of argument: the other side is heard at its best before verdict. This both protects minorities and restrains majorities from epistemic domination.

Mitigating asymmetries. Power, expertise, and rhetorical skill are unevenly distributed. Steelmanning partially offsets these asymmetries by:

- Role-reversal fairness: the stronger party must articulate the weaker party's case credibly.
- Inclusion of standpoint knowledge: the practice makes room for situated insights without demanding prior mastery of elite idioms.
- Guardrails against harm: when restatement itself risks amplification of harm, a minimal, documented version with explicit ethical justification preserves process without complicity.

Civic dividend. Habits of steelmanning increase trustworthiness of institutions (journals, courts, councils) by making reasoning auditable. Citizens learn to distinguish defeat by fairness from suppression by force.

3.3 Virtue-epistemic: humility, courage, generosity, and integrity

Humility (teachability). Steelmanning confesses fallibility in practice: *my opponent may be tracking a truth I have missed.* By seeking the strongest version, I place myself under the discipline of possible correction—the core of intellectual humility.

Courage (risking movement). To strengthen another's argument is to risk being moved by it. Courage here is not bravado but willingness to face dissonance without defensive distortion—bracketing status incentives to serve truth.

Generosity (charity with spine). Generosity supplies missing but admissible premises, removes inessential weak points, and names the good the view seeks (security, fairness, freedom, dignity). It refuses the cheap wins of misrepresentation while still holding firm lines against error and harm.

Integrity (truthfulness in representation). Integrity binds us to say what is so, as best we can tell—including when "what is so" helps an opponent. Misstating another's position is a corruption of inquiry; steelmanning is its cure. Integrity also requires scope discipline: neither inflating a claim beyond its warrant nor shrinking it to something vacuous.

Virtue ecology. Practiced together, these virtues create an inquiry culture in which reasons can meet: humility opens ears; generosity opens space; courage opens risk; integrity keeps the rails straight. The effect is cumulative: over time, adversaries are transformed into co-inquirers, and disagreement becomes a shared project rather than a zero-sum contest.

4. Method: The Philosophical Steelman

4.1 Interpretive charity and the principle of repair (Davidson)

Begin from the Davidsonian maxim: interpret others so that most of what they say comes out true or reasonable by their own lights. Charity is not flattery; it is the minimal condition for interpretation to succeed at all. The principle of repair follows: when a text or utterance appears inconsistent or trivially false, first try local repairs that preserve intent (clarify an equivocation, supply an implicit restriction, correct an obvious slip). Only when reasonable repairs fail should we attribute deep error. Repair is bounded by fidelity: do not introduce theses the speaker would reject, nor delete commitments central to their stance.

Example. If someone says, "Science is only a social construct," charity hears a critique of naive scientism, not a denial that planes fly. Repair reads it as a claim about underdetermination and power, then tests that stronger, coherent version.

4.2 Best-case reconstruction: supplying admissible premises, pruning inessentials

A best-case reconstruction presents the argument with (i) logical structure made explicit, (ii) the strongest premises the proponent would accept, and (iii) inessential or weak add-ons removed. "Admissible" means: the added premise is either stated elsewhere by the author, entailed by

their commitments, or would be endorsed on reflection. “Pruning” means: drop rhetorical excess that does no real work, and trim scope to the domain where the view has its best chance.

Norms.

- Map the argument as a numbered list of premises and a conclusion.
- Replace vague quantifiers with the strongest defensible restriction (e.g., “often,” “in domains D”).
- Cite or quote where each premise is held or naturally entailed.

Pitfalls. Do not smuggle in premises they reject to make the view easier to defeat. Do not shrink the claim into a platitude that evades genuine evaluation.

4.3 Telos disclosure: naming the good the view aims to secure

Every serious position presses toward some good (telos): justice, dignity, security, liberty, truth, solidarity, sanctity, stewardship, peace of conscience, etc. Telos disclosure requires naming that good explicitly. Doing so reveals why reasonable people might hold the view, clarifies tradeoffs, and makes convergence possible even when conclusions differ. It also blocks bad-faith portrayals that reduce opponents to mere error or malice.

Example. A free speech maximalist seeks protection against coercive silencing that chills inquiry; a content-moderation advocate seeks protection against harms that corrode dignity and participation. Naming both goods focuses the disagreement on mechanisms and thresholds, not caricatures.

4.4 Public-reason rendering: reasons legible to non-adherents

In plural settings, arguments should be rendered in terms that those who do not share one’s metaphysical or sectarian commitments can nonetheless evaluate. Public-reason rendering does not forbid comprehensive doctrines; it asks for bridgeable reasons: empirical claims with sources, principles that track widely shared values, and procedures that admit common standards of assessment. The aim is legibility and reciprocity, not self-censorship.

Practices.

- Translate sectarian premises into publicly testable surrogates where possible (e.g., from “sacredness of life” to “protection of persons and vulnerability”).

- Distinguish what follows from shared standards vs. what rests on tradition-specific axioms.
- Offer conditional forms: “If we accept standards S, then policy P follows.”

4.5 The satisfaction test: endorsement by the interlocutor

A reconstruction is a steelman only if it passes the satisfaction test: the interlocutor, given the paragraph you wrote, says, “Yes, that is my view, and stated as well as or better than I could.” Treat this as a procedural norm, not a courtesy. Seek explicit assent; invite edits; incorporate them faithfully. Once endorsed, the steelman becomes the target of evaluation. This prevents moving goalposts, reduces defensiveness, and sharpens the subsequent inquiry.

Operational steps.

1. Read back the reconstruction neutrally.
2. Ask for a 1–10 closeness score and concrete edits.
3. Version and timestamp the agreed text.
4. Proceed only once endorsement is explicit.

Why it matters. Without endorsement, one risks refuting a position no one holds. With endorsement, critique is constrained to the genuine crux, and disagreement can become a joint search for the cheapest decisive test.

5. Hard Cases and Limits

5.1 Performative self-contradictions and minimal restatement

Some utterances undermine the very norms they invoke—they are *performatively self-contradictory*. Examples:

- “No statements are truth-apt.” (presented as a truth-apt statement)
- “Reasons are tools of oppression; therefore you should accept mine.”
- “Debate is illegitimate,” offered *as* a move in debate.

Norm. Apply minimal restatement: register the claim at the level needed to expose the contradiction, do not fortify it beyond coherence, and justify the limit. You may write:

“The position rejects reasons as binding while offering reasons to bind. Because the act contradicts the content, further strengthening would misdescribe it. We therefore restate minimally and assess on those grounds.”

Why this is not evasion. Steelmanning presumes a floor of rational cooperativity (Gricean maxims, non-contradiction). When a claim denies that floor, repair must stop where coherence ends. Minimal restatement preserves fidelity without laundering incoherence.

5.2 Harm-amplifying speech and justified refusal to steelman

Some claims, even when coherently expressible, predictably amplify harm (e.g., targeted dehumanization, incitement, doxxing instructions). Full-strength reconstruction may provide a sharper rhetorical weapon.

Principle. The duty to steelman is bounded by the duty to avoid foreseeable, unjustified harm. In such cases, one may refuse full steelmanning and instead:

1. Identify the core proposition neutrally (to avoid the hollow-man fallacy).
2. State the risk rationale for not fortifying (audience, context, history of effects).
3. Offer scope-limited assessment: address the claim’s best-known arguments at a general level or within safe hypotheticals.
4. Propose alternative venues (closed review, expert committee) if rigorous testing is warranted.

Template.

“We decline to strengthen this formulation because amplification risks concrete harm (reasons: …). We record the core claim as X, answer it at a general level, and invite vetted, non-public review if further analysis is needed.”

Guardrails against abuse. The refusal must be reasoned, reviewable, and narrow, lest “harm” become a pretext for suppressing mere discomfort.

5.3 Radical incommensurability: translation before adjudication

Disputants sometimes reason within distinct language-games—different basic concepts, evidential standards, or value-weightings—so that direct evaluation falls flat. Examples: econometric cost–benefit vs. rights-based deontic claims; experimentalism vs. revelation; mechanistic explanation vs. teleological interpretation.

Method: translation first.

- Term mapping: identify overlaps (shared concerns) and false friends (same word, different use).
- Standard disclosure: specify what counts as evidence or warrant *within each frame*.
- Conditional bridges: "Given standards S1, claim C has weight W; given S2, weight W'."

Outcomes.

- If bridges suffice, proceed to adjudication on common ground.
- If not, shift to metadisciplinary negotiation: which standards are appropriate for *this* decision-context (law, policy, pedagogy, research)?
- Failing that, seek modus vivendi: carve domains where each standard governs without mutual veto.

Limit. Steelmanning cannot conjure a single tribunal when none exists; it can clarify where and why standards diverge and prevent category mistakes.

5.4 Motte-and-bailey dynamics and boundary-setting

A motte-and-bailey strategy alternates between a bold, controversial thesis (bailey) and a defensible, modest claim (motte). When pressed, the proponent retreats to the motte; when safe, they advance the bailey—trading on their confusion.

Steelmanning without abetting equivocation.

1. Dual reconstruction: explicitly state both versions.
    - *Motte:* the narrow, defensible claim (with scope and conditions).
    - *Bailey:* the expansive claim actually used rhetorically.
2. Nonporous boundaries: mark where evidence for the motte does not license the bailey.
3. Burden of bridge: require the proponent to provide valid bridging arguments (analogy, generalization, mechanism) or to abandon the bailey.
4. Operational constraint: agree to discuss only one version at a time; tabling the other prevents slippage.

Formula (usable in text).

“For clarity, we distinguish Motte (M) and Bailey (B). We steelman M as follows: … We separately state B: … Evidence E supports M within domain D; E does not, without additional argument A, support B. We invite A; absent it, evaluation proceeds on M alone.”

Payoff. Boundary-setting preserves the virtue of charity without becoming a conduit for equivocation. It also teaches good argumentative hygiene: scope discipline is part of truthfulness.

Summary of limits. Steelmanning presumes *coherence, cooperativity,* and *non-harm*. When these are missing or compromised, the method adapts: minimal restatement for incoherence, justified refusal or indirection for harm, translation for incommensurability, and boundary-setting for motte-and-bailey. The goal remains constant: protect the conditions under which reasons can meet without distortion or injury.

6. Exemplars (Textual Miniatures)

6.1 Aquinas: strongest objections first, then answer

Form. *Objectiones → Sed contra → Respondeo → Ad primum…* (replies). Aquinas elevates the opponent’s case—often tightening it beyond its sources—then answers without scorning what is true in it.

Miniature (modelled on the form).
Q: Whether public dissent ought to be tolerated in a just polity.

Objection 1. It should not, because dissent sows discord; unity is a common good, and whatever destroys the common good is to be prohibited.
Objection 2. Further, falsehood harms souls; to tolerate its proclamation is to consent to harm.
Objection 3. Further, authorities are ordained to uphold truth; toleration implies indifference to their charge.

Sed contra. Augustine: “Many things are tolerated for the sake of peace.”

Respondeo. I answer that toleration is not approval but a prudential judgment where (a) coercion would cause greater evils (hypocrisy, rebellion, epistemic stagnation), and (b) the common pursuit of truth is better served by contestation under just limits. Since human judgment is fallible and truth in public matters is often discovered dialectically, a polity should

permit dissent *as a rule*, restraining only speech that directly and imminently violates justice (e.g., targeted incitement). Thus toleration preserves peace and the path of correction.

Ad 1. Unity without truth is brittle; ordered pluralism better secures the common good.
Ad 2. The remedy for falsehood in open questions is counter-speech and education, not prior restraint.
Ad 3. Authority better upholds truth by hosting reasons than by foreclosing them.

Lesson. Steelmanning here means: pose the most stringent anti-toleration case first, then answer by granting its proper goods (peace, truth) and showing a narrower path that secures them more reliably.

6.2 Mill: live adversaries vs. "dead dogma"

Core insight. Beliefs held without confrontation decay into "dead dogma." A view remains alive only if it meets its best opponents under fair conditions.

Miniature (Millian key).
Take the claim: "Compelled consensus secures social harmony." The Millian steelman grants the good (harmony), then argues: harmony purchased by silencing dissent rots its own warrant—errors persist uncorrected; truths become slogans detached from reasons; minorities are taught that legitimacy lies in power, not argument. The cure is not cacophony but institutions that expose beliefs to their strongest critics (free press, adversarial review, open classrooms). Even true doctrines need adversaries; without them, we forget *why* they are true and lose the capacity to defend or improve them.

Lesson. Steelmanning, in Mill's register, keeps doctrines living by insisting that they face the best case against them—or else forfeit their claim on rational allegiance.

6.3 Dennett/Rapoport: endorsement-standard restatement

Rule-set. (1) State the opponent's view so well they say, "Thanks—better than I put it." (2) List agreements. (3) Note what you learned. (4) Only then, critique.

Miniature (applied to a contemporary dispute).
Topic: "Strong encryption should be legally unbreakable."

Restatement (aiming for endorsement).
You argue that general-purpose encryption must remain unbreakable by design, because (a) secure private communication is a precondition for autonomy, commerce, and dissent; (b)

"exceptional access" creates systemic vulnerabilities that attackers inevitably exploit; (c) history shows that narrow backdoors expand in scope and chill speech; and (d) law enforcement can pursue targeted ends through endpoints and warrants, without weakening the cryptosystem for everyone. Your aim is not to shield crime but to preserve a public good—a trustworthy digital commons.

Agreements. We agree that (i) privacy and security are mutually reinforcing at the systems level; (ii) exceptional-access proposals have repeatedly leaked or been abused; (iii) endpoint-focused warrants are often more proportionate.

What I learned. Your point about "function creep" shows how even well-scoped access morphs under institutional pressures.

Critique (now). Where we differ is emergency response: in narrowly defined, judicially supervised crises (e.g., imminent mass-casualty threats), I argue for time-bound, hardware-specific escrow with public auditing. The crux is empirical: can such schemes avoid systemic risk? Let's propose a cheapest decisive test—a pilot with red-team oversight and automatic sunset—while keeping your core good (no weakened general crypto) intact.

Lesson. The endorsement standard converts sparring into co-design: after the opponent says "Yes, that's mine," common aims surface, cruxes sharpen, and proposals become testable.

7. Applications

7.1 Classroom disputation and seminar practice

Pedagogical aim. Replace performative debate with co-inquiry. Students learn to *understand before evaluating*, internalizing accuracy, charity, and scope discipline.

Practices.

- Two-step turns. Any critique must be preceded by a steelman restatement; instructor cold-calls the *opponent* to score its fidelity (1–10) before discussion proceeds.
- Rotating roles. Assign "Author," "Steelman," and "Crux-finder." The crux-finder's task is to identify the narrowest hinge on which disagreement turns (a premise, a method, or a value priority).
- Textual miniatures. Begin sessions with a 150–200 word Aquinas-style miniature: *Objectiones* (two strongest), *Respondeo* (answer granting goods), *Ad* (targeted replies).
- Rubric. Grade for (i) fidelity (no invention/omission), (ii) best-case strengthening, (iii) telos disclosure, (iv) scope/limits named, (v) clarity of crux.

- Assessment. Replace “gotcha” exams with paired reconstructions: each student must secure the interlocutor’s written endorsement of a one-paragraph steelman.

Payoff. Higher signal discussions; reduced defensiveness; measurable gains in conceptual clarity and civility.

7.2 Editorial/journal peer review norms

Editorial aim. Make fairness auditable and reduce Type-I refutation errors.

Practices.

- Steelmanning paragraph required. Reviewers open with a 120–180 word Best-Case Summary. Authors rate its fidelity; editors see the rating.
- Objection formatting. Reviews separate (A) strongest concession-worthy merits from (B) decisive objections; the latter target the agreed steelman, not a straw version.
- Public-reason rendering. For interdisciplinary submissions, reviewers state how claims are legible beyond the subfield (methods, standards of evidence, limitations).
- Revision targets. Decision letters anchor required changes to the agreed steelman (e.g., “Widen the scope statement; add counterexample X; report uncertainty Y”).
- Appeal pathway. Authors may appeal *only* on grounds of misrepresentation of the steelman; editors adjudicate by comparing the author-endorsed summary and the review.

Payoff. Cleaner correspondence between critique and claim; fewer needless rejections; better cross-field uptake.

7.3 Civic forums and policy deliberation (public reason)

Civic aim. Convert adversarial hearings into reciprocal justification fit for plural publics.

Practices.

- Paired presentations. Each stakeholder must present a steelman of the opposing brief before presenting their own. Chairs suspend testimony that fails the minimal fidelity test.

- Claim + Telos cards. Policy memos begin with (i) Claim (one sentence), (ii) Evidence/Method, (iii) Uncertainty, (iv) Telos (goods protected), (v) Scope/limits. Opponents must accept the card's accuracy before rebutting.
- Crux calendar. After first session, moderators publish agreed cruxes (e.g., data reliability; threshold of harm; rights tradeoff). Subsequent sessions are organized by crux, not by faction.
- Cheapest decisive test. Where possible, parties co-design pilot programs or sunset clauses that test the disputed hinge while minimizing risk.
- Harm guardrail. For topics with plausible harm amplification (e.g., targeted misinformation), forums adopt justified partial steelmanning: neutral core statement, no rhetorical strengthening, with reasoned explanation.

Payoff. Fewer talking-past-each-other exchanges; policies shaped around real hinges; increased perceived legitimacy.

7.4 Interfaith, science–religion, and identity-implicating debates

Dialogical aim. Maintain fidelity and dignity across deep differences while preserving evaluability.

Practices.

- Dual-register steelman. Participants offer (A) an internal restatement (using the tradition's own terms) and (B) a public-reason rendering (bridgeable to non-adherents). Both require endorsement by the originator.
- Telos pairing. Each side explicitly names the good it seeks (sanctity, dignity, truth, justice, solidarity, stewardship, freedom of conscience). Moderators track overlaps and tradeoffs.
- Translation before adjudication. Where standards diverge (revelation vs. experiment; rights vs. utility), sessions begin with method disclosure: what counts as evidence or warrant within each frame; then proceed by conditional arguments ("If we accept S, then P follows").
- Minimal restatement & refusal protocols. For claims that are performatively self-contradictory or plausibly harm-amplifying, adopt minimal restatement or justified refusal (documented, narrow, reviewable).

- Story + reason pairing. Permit narrative witness (lived experience) alongside argument, but require a companion reason card that states claims and limits legibly to outsiders.

Payoff. Disagreement without contempt; clearer boundaries; occasional genuine convergence; when not, principled modus vivendi instead of stalemate.

Summary of application ethos. In each venue, steelmanning is not a nicety but infrastructure: a public, repeatable discipline that (i) secures fairness, (ii) sharpens cruxes, and (iii) yields actions or revisions that the other side can recognize as responsive—even when the verdict is "we still disagree."

8. Objections and Replies

8.1 "Charity enables bad ideas" → filtering by harm/performative incoherence

Objection. By strengthening objectionable views you risk laundering them—granting fringe or dangerous claims undue legitimacy and rhetorical polish.

Diagnosis. The worry conflates charitable interpretation with normative endorsement. It also ignores two well-marked limits: (i) performative self-contradiction (claims that undercut the very norms of reason they invoke), and (ii) foreseeable harm amplification (restatements that predictably facilitate injury—e.g., targeted dehumanization, doxxing).

Reply. Steelmanning is bounded by fidelity, coherence, and non-harm. When a claim is incoherent, employ minimal restatement: expose the contradiction without fortifying it. When amplification risks unjust harm, a justified partial or refused steelman is permissible: neutrally record the core proposition, state the risk rationale, and answer at a general level or in vetted venues. Charity need not—and must not—become complicity.

Policy (actionable).

- Screen 1: Does the claim meet baseline coherence? If not, minimal restatement.
- Screen 2: Would fortification foreseeably amplify harm? If yes, document a narrow refusal and proceed with a non-amplifying critique.
- Screen 3: Otherwise, steelman and evaluate the *best admissible* version.

Illustrative contrast.

- *Permissible steelman:* "Abolish qualified immunity to enhance police accountability" (coherent; public reasons available).

- *Refused fortification:* “Out this private individual; they are subhuman” (harm-amplifying; fails public-reason standard).

8.2 “It stacks the deck for opponents” → reciprocity and symmetry

Objection. Strengthening the other side disadvantages your own view, handing them free improvements you must then overcome.

Diagnosis. This frames inquiry as a zero-sum contest of status. In truth-seeking contexts, asymmetric fairness is unstable: if only one side steelmans, the other enjoys cheap wins via straw men; if neither steelmans, discourse degrades into noise.

Reply. Steelmanning is reciprocal and symmetric by norm. Each side owes the other an endorsement-standard restatement; editors, moderators, and teachers can enforce this symmetry. Far from “stacking the deck,” the practice equalizes it: both positions face their best versions; cheap victories disappear; only real advantages (better evidence, tighter inference, clearer value trade-offs) remain. If your view cannot withstand its opponent’s best case, the problem is not the method but the view.

Policy (actionable).

- Require mutual steelman paragraphs (120–180 words) with explicit endorsement before critique.
- Separate merits from objections in reviews or debates; target the agreed steelman only.
- Publish crux lists so that each side’s strongest hinge is addressed, not the weakest outskirts.

Result. Reciprocity turns “giving the opponent a ladder” into building a shared scaffold. If someone refuses reciprocity, that refusal itself is evidential about their confidence and fairness.

8.3 “It’s inefficient” → crux-finding and cheapest decisive tests

Objection. Steelmanning takes time; in classrooms, journals, and policy rooms, we cannot afford elaborate reconstructions before getting to the point.

Diagnosis. The alternative—serial skirmishes with misread positions—*feels* faster but wastes cycles on irrelevancies, entrenches defensiveness, and delays resolution. Apparent speed hides downstream inefficiency.

Reply. A short, front-loaded steelman reduces total cost by accelerating crux-finding—the identification of the narrow hinge on which disagreement turns. Once the crux is explicit, disputants can design the cheapest decisive test (CDT): the smallest observation, pilot, or conditional analysis likely to move either side. Time invested in faithful reconstruction is repaid by fewer iterations, clearer decisions, and lower error rates.

Policy (actionable).

- Time-box the practice: two minutes for a steelman read-back; one minute for endorsement/edits.
- Institutionalize Claim + Crux cards (one sentence each) appended to the steelman.
- Tie decisions to CDTs: pilots, adversarial red-team checks, or pre-registered analyses with sunset clauses.

Empirical analogue. In software and science, brief spec clarifications save orders of magnitude of rework. Steelmanning is the discursive equivalent: a small spec of the opponent's claim prevents costly misbuilds of the argument one intends to test.

UpShot. Properly bounded and reciprocally enforced, steelmanning filters bad ideas rather than laundering them, levels the field rather than tilting it, and saves time by cutting to the crux and proposing the cheapest decisive test.

9. A Compact Procedure (Philosophy-First)

9.1 Reconstruct → Confirm → Identify crux → Propose test → Answer

Step 1 — Reconstruct (faithful + best-case).
State the interlocutor's view in its *strongest, admissible* form: clarify terms, supply only premises they would endorse, prune inessentials, declare scope and limits, and name the good (telos) the view seeks.

Step 2 — Confirm (endorsement on record).
Invite explicit assent. If they revise, adopt their edits. Version the agreed paragraph (time/date). Without endorsement, you risk refuting a view no one holds.

Step 3 — Identify the crux (the narrow hinge).
Locate the one claim/assumption whose change would flip at least one party's conclusion: a contested fact, evidential standard, model choice, or value priority. Write it in one sentence.

Step 4 — Propose the cheapest decisive test (CDT).
Offer the smallest observation, analysis, pilot, or conditional argument likely to move either side on the crux, with clear pass/fail criteria, costs, and sunset/replication plans.

Step 5 — Answer (reasoned stance under conditions).
Respond to the agreed steelman: grant what follows if the CDT favors them; state what follows if it favors you; give your present verdict with reasons, uncertainties, and any ethical guardrails.

9.2 Exemplar paragraph template and satisfaction criterion

A. One-paragraph Steelman (fill-in template)

Claim. C: [one sentence].
Definitions/scope. By "[term]" you mean [definition]; the claim concerns domain D and excludes [edge cases].
Best reasons. (1) [strongest reason], (2) [next strongest], supported by [most probative evidence or mechanism].
Telos. The view aims chiefly to secure [good(s): e.g., fairness, dignity, safety, truth].
Acknowledged limits. The claim holds under [conditions] and is vulnerable when [known weakness].
UpShot. On this reading, the position is: [succinct restatement integrating the above].

Satisfaction criterion (endorsement test).

- Ask: *"If this paragraph appeared under your name, would you endorse it without caveat?"*
- Record: closeness score (1–10) + requested edits.
- Only proceed once the paragraph is endorsed or the refusal is documented (with reasons).

B. Crux + CDT card (one sentence each)

- Crux. *Our disagreement turns on whether* [precise hinge].
- Cheapest decisive test. *We can resolve movement on the crux by* [pilot/analysis/observation], *which counts as decisive if* [criteria] *by* [date/sunset].

C. Answer stub (conditioned conclusion)

- If CDT outcome = A: I will accept [implication], revise [policy/claim], and note [remaining caveat].

- If CDT outcome = ¬A: I maintain [implication], propose [next step], and commit to [monitoring/replication].
- Current verdict: Given present evidence, I hold [position] because [reasons], while granting [concessions] and respecting [ethical guardrails].

Worked micro-example (contemporary, neutral)

Steelman (endorsed).
Claim. The city should implement congestion pricing downtown.
Definitions/scope. “Congestion pricing” = variable road tolls within zone Z during peak hours; applies to private vehicles, not emergency services or disability exemptions.
Best reasons. (1) Pricing reduces peak traffic volume and delays by shifting marginal trips; (2) revenue funds transit that benefits lower-income commuters; evidence from London/Stockholm shows sustained reductions in travel times and emissions.
Telos. Efficiency, air quality, equitable mobility.
Acknowledged limits. Risk of burdening small businesses and workers with inflexible hours; effectiveness depends on viable transit alternatives.
UpShot. Adopt time-varying tolls in Z, paired with targeted transit improvements and exemptions.

Crux. Whether comparable cities’ effects will generalize *given our current transit capacity*.
CDT. Six-month pilot: peak-hour pricing in Z with pre-registered metrics (avg. travel time, bus speed, PM2.5, small-business receipts). Decisive if: bus speeds +15% and PM2.5 −10% with no >2% revenue loss for a matched basket of small firms (net of mitigation).

Answer (now). I provisionally support the pilot conditional on mitigation credits for shift workers and real-time audits. If metrics fail thresholds, sunset and revert; if they pass, scale with published equity adjustments.

Bottom line. The compact procedure keeps philosophy first—fidelity in reconstruction, clarity at the crux, frugality in testing, and integrity in answering—so that disagreement becomes jointly managed inquiry rather than a contest of misfires.

10. Conclusion: From Victory to Understanding

10.1 Steelmanning as discipline of shared reason

To “win” without having represented the other side faithfully is a kind of epistemic theft: the surplus comes from misdescription, not merit. Steelmanning reverses the incentive. It binds disputants to a prior oath of fairness: I will not advance my case against a position you would not sign. That oath does three quiet things. First, it stabilizes meaning—terms are fixed, scope is declared, hedges are honored. Second, it clarifies stakes—the good a view seeks (its telos) is named, so objections must answer to that good rather than a caricature. Third, it enables correction—once the strongest admissible form is public, error can be owned without humiliation, because the process already dignified the opponent. In this sense, steelmanning is not a conversational courtesy but a discipline of shared reason: it restores the conditions under which reasons can actually meet, travel, and change minds.

10.2 Cultivating persons and publics capable of disagreement without contempt

Contempt is a shortcut: it spares us the labor of understanding. Steelmanning reintroduces that labor and turns it into formation. At the personal level, it trains (i) humility—I begin by learning to say your view as well as you do; (ii) courage—I risk movement by strengthening what could defeat me; (iii) generosity—I supply missing but admissible premises; (iv) integrity—I refuse the gains of misrepresentation. At the public level, the practice scales through small, enforceable norms: classrooms that grade the fidelity of restatements; journals that require author-endorsed best-case summaries; councils that structure hearings around crux lists rather than faction slots; broadcasters and moderators who pause segments until each side can pass a one-paragraph endorsement test. When esteem accrues to fidelity and crux-finding rather than to volume and derision, a culture becomes capable of disagreement without contempt—the form of conflict democracy can survive.

10.3 Future work: metrics for discourse quality; pedagogy and institutional uptake

Moving from virtue to institution requires measurement and adoption.

- Metrics (make fairness auditable):
    - *Fidelity index:* share of debates/reviews where the target endorses the restatement (goal: >80%).
    - *Misrepresentation rate:* frequency of post hoc corrections for straw/hollow man (goal: <5%).
    - *Crux efficiency:* median time (or pages) to an agreed crux; fraction of talk tethered to that hinge.

    - *Decision clarity:* presence of a cheapest decisive test (CDT), pre-registered thresholds, and an explicit sunset or update rule.
    - *Correction half-life:* median time from decisive counterevidence to public revision or retraction.

- Pedagogy (from habit to curriculum):
    - Canon + craft: pair Platonic/Aquinian exemplars with live steelman practicums.
    - Assessment: require endorsement-standard paragraphs; grade for telos disclosure, scope discipline, and crux articulation.
    - Capstones: students co-design a CDT for a live controversy and publish a brief post-mortem at D+30.
- Institutional uptake (simple switches with large effects):
    - Journals: mandate a 150-word steelman at the top of every review; editors see the author's fidelity score.
    - Policy fora: open hearings with paired steelmen; deliberations proceed by crux, not by turn-taking.
    - Media & platforms: add a "Pass the Paragraph" badge when both sides endorse each other's summaries; demote segments that fail fidelity checks.
    - Courts & committees: weave steelman clauses into rules of procedure—no ruling or recommendation without an endorsed statement of the losing view.

None of this requires agreement about ultimate grounds. It requires a modest civic covenant: before we disagree, we will understand; before we judge, we will state the other's view as they themselves would. Under that covenant, argument recovers its vocation—not victory over an enemy, but understanding with an adversary, and, sometimes, genuine convergence.

References

- Aquinas, Thomas. *Summa Theologiae*. Latin text and English trans. by the Fathers of the English Dominican Province. London: Burns Oates & Washbourne, 1912–36 (repr. many eds.).
- Aristotle. *Topics*, *Rhetoric*, and *Nicomachean Ethics*. In *The Complete Works of Aristotle*, ed. Jonathan Barnes. Princeton: Princeton University Press, 1984.
- Aumann, Robert J. "Agreeing to Disagree." *The Annals of Statistics* 4, no. 6 (1976): 1236–39.
- Davidson, Donald. "Radical Interpretation." *Dialectica* 27, nos. 3–4 (1973): 313–28.
- Dennett, Daniel C. *Intuition Pumps and Other Tools for Thinking*. New York: W. W. Norton, 2013. (Ch. "Rapoport's Rules.")
- Gadamer, Hans-Georg. *Truth and Method*. 2nd rev. ed. Trans. Joel Weinsheimer and Donald G. Marshall. New York: Continuum, 2004 (orig. 1960).
- Grice, H. P. "Logic and Conversation." In *Syntax and Semantics, Vol. 3: Speech Acts*, ed. Peter Cole and Jerry L. Morgan, 41–58. New York: Academic Press, 1975.
- Habermas, Jürgen. *The Theory of Communicative Action*. 2 vols. Trans. Thomas McCarthy. Boston: Beacon Press, 1984–87 (orig. 1981).
  ——. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Trans. William Rehg. Cambridge, MA: MIT Press, 1996 (orig. 1992).
- Mill, John Stuart. *On Liberty*. London: John W. Parker and Son, 1859.
- Quine, W. V. O. *Word and Object*. Cambridge, MA: MIT Press, 1960.
- Rawls, John. *Political Liberalism*. Expanded ed. New York: Columbia University Press, 2005 (orig. 1993).
  ——. "The Idea of Public Reason Revisited." *University of Chicago Law Review* 64, no. 3 (1997): 765–807.
- Rapoport, Anatol. *Fights, Games, and Debates*. Ann Arbor: University of Michigan Press, 1960.
  ——. *Strategy and Conscience*. New York: Harper & Row, 1964.
- Roberts, Robert C., and W. Jay Wood. *Intellectual Virtues: An Essay in Regulative Epistemology*. Oxford: Oxford University Press, 2007.

- van Eemeren, Frans H., and Rob Grootendorst. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge: Cambridge University Press, 2004. (For norms of critical discussion; straw/hollow-man diagnostics.)
- Walton, Douglas. *Informal Logic: A Pragmatic Approach*. 2nd ed. Cambridge: Cambridge University Press, 2008. (Chs. on straw man and dialogue shifts.)
- Zagzebski, Linda Trinkaus. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press, 1996.
- Plato. *Republic*, *Gorgias*, *Meno*. In *Plato: Complete Works*, ed. John M. Cooper. Indianapolis: Hackett, 1997.

Optional editions/translations cited in teaching (helpful for page references)

- Aquinas. *Summa Theologiae*, Latin–English (Blackfriars edition), 61 vols. London: Eyre & Spottiswoode; New York: McGraw-Hill, 1964–81.
- Aristotle. *Rhetoric*. Trans. George A. Kennedy. Oxford: Oxford University Press, 2007.
- Mill. *On Liberty and Other Writings*. Ed. Stefan Collini. Cambridge: Cambridge University Press, 1989.
- Gadamer. *Truth and Method*, rev. 2nd English ed. London: Bloomsbury, 2013.
- Davidson. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, 1984.